
DYNANEWTON

Accelerating Newton’s Method for Machine Learning

Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann
 Department of Computer Science, ETH Zurich, Switzerland

Abstract

Newton’s method is a fundamental technique in optimization with quadratic convergence within a neighborhood around the optimum. However reaching this neighborhood is often slow and dominates the computational costs. We exploit two properties specific to empirical risk minimization problems to accelerate Newton’s method, namely, subsampling training data and increasing strong convexity through regularization. We propose a novel continuation method, where we define a family of objectives over increasing sample sizes and with decreasing regularization strength. Solutions on this path are tracked such that the minimizer of the previous objective is guaranteed to be within the quadratic convergence region of the next objective to be optimized. Thereby every Newton iteration is guaranteed to achieve super-linear contractions with regard to the chosen objective, which becomes a moving target. We provide a theoretical analysis that motivates our algorithm, called DYNANEWTON, and characterizes its speed of convergence. Experiments on a wide range of data sets and problems consistently confirm the predicted computational savings.

1 Introduction

In machine learning, we often fix a function class with parameters $\mathbf{x} \in \mathbb{R}^d$, define a non-negative family of loss functions ϕ^z , a regularizer Ω , and then aim to minimize a regularized sample loss over training data \mathcal{S} ,

$$f_{\nu}^{\mathcal{S}}(\mathbf{x}) := \frac{1}{|\mathcal{S}|} \sum_{z \in \mathcal{S}} \phi_{\nu}^z(\mathbf{x}), \quad \phi_{\nu}^z(\mathbf{x}) := \phi^z(\mathbf{x}) + \nu \Omega(\mathbf{x}). \quad (1)$$

Justified by theories like Tikhonov regularization or structural risk minimization, we know that we can control the expected risk of the minimizers \mathbf{x}_{ν}^* of $f_{\nu}^{\mathcal{S}}$, i.e. avoid overfitting, by choosing the regularization strength ν appropriately.

In this paper, we focus on Newton’s method for optimization, which obeys quadratic convergence towards an extremal point, when initialized sufficiently close to such a point [20]. However, despite this unmatched speed of convergence, Newton’s method has shortcomings for large-scale machine learning problems of the type described above: (i) Reaching the quadratic convergent regime through an initial *damped phase* [7] is where most of the computation is typically spent, unless one has access to an initial solution close-enough to the optimum. (ii) Being a batch algorithm, each Newton iteration involves a complete pass over the entire data set to compute the required gradient and Hessian matrix. (iii) The computation of the Newton update requires to solve a linear system of equations, which is a challenge, in particular, if the data dimensionality is high.

The strategy presented in this paper is as follows. Following the ideas of [8], we present a systematic way to dynamically subsample the data so as to match-up statistical and computational accuracy, leading to significant savings in the amount of overall computation. Moreover, employing this together with an adaptive control of the regularization strength, we define a *continuation method* [1],

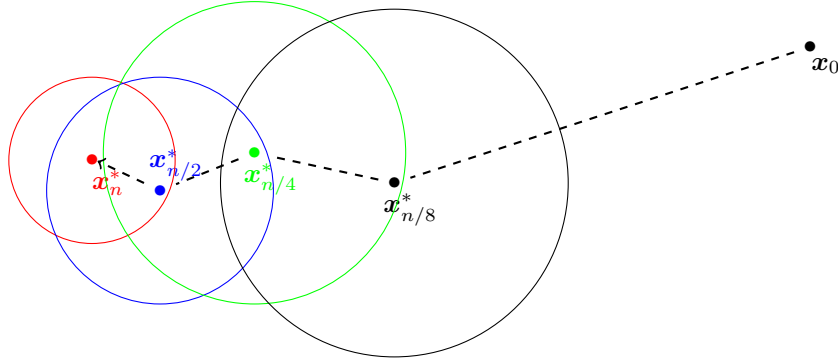


Figure 1: Illustration of the continuation method approach proposed in this paper: Assume we chose $\nu \propto 1/n$. After optimizing the risk for a small subsample, we continuously increase the sample size (by a factor, e.g. 2) and proportionally lower ν . We want to guarantee that each solution provides a starting point that is within the quadratic convergence region of the subsequent optimization. These regions are depicted by colored circles.

where we track solutions computed for problems with fewer data and with stronger regularization. We use previous solutions in order to compute the starting point for the next Newton iteration(s), possibly operating on a larger sample. An ideal sketch of the situation is shown in Figure 1. This is meant to address challenges (i) and (ii). Finally, in order to overcome challenge (iii), we also empirically investigate our strategy for a popular quasi-Newton method, namely BFGS, which computes approximations to the inverse Hessian through closed-form rank-one matrix updates.

2 Related work

Adaptive data sub-sampling Recent results have shown that the finite sum structure of the empirical risk can be exploited to achieve linear convergence for strongly convex objectives [15, 10, 23]. This is accomplished by revisiting samples and storing their gradients in order to reduce the variance of future update directions. Although these methods achieve fast convergence on the empirical risk, they do not explicitly consider the expected risk, which has been separately studied in the literature on learning theory. It is usually analysed with the help of uniform convergence bounds that take the generic form [5]

$$\mathbf{E}_{\mathcal{S}} \left[\sup_{\mathbf{x} \in \mathcal{F}} |f^{\mathcal{S}}(\mathbf{x}) - \mathbf{E}_{\mathbf{z}} \phi^{\mathbf{z}}(\mathbf{x})| \right] \leq \mathcal{H}(n), \quad (2)$$

where the expectation is over a random sample \mathcal{S} of size n . Here \mathcal{H} is a bound that depends on n , usually through a ratio n/d , where d is the capacity of \mathcal{F} (e.g. VC dimension) [6].¹

The recent work of [8] simultaneously exploits the properties of the empirical risk as well as the concentration bounds from learning theory to achieve fast convergence to the expected risk. This approach uses a dynamic sample size schedule that matches up optimization error and statistical accuracy. Although this approach was tailored specifically for variance-reduced stochastic gradient methods, we here show how a similar adaptive sample size strategy can be used in the context of non-stochastic approaches such as Newton’s method.

Regularization paths and continuation methods There is a rich body of literature on numerical continuation methods, a reference work being [1]. The basic idea is to define a family of objectives, in the simplest case with a single parameter $t \in [0; T]$, and to optimize over a sequence of objectives f_t with increasing t , such that the sequence approaches some desired final $f = f_T$. The general motivation is that following the solution path $\mathbf{x}_t^* = \arg \min_{\mathbf{x}} f_t(\mathbf{x})$ may be computationally more efficient than optimizing the (typically) harder problem f directly.

¹For ease of exposition, we assume that all regularized risks f_{ν} can be governed by the same function \mathcal{H} .

Free energy based continuation methods, often for non-convex or integer problems, have been popularized in computer vision and machine learning under the name of *deterministic annealing* [21], a deterministic variant of simulated annealing [16]. Here the family of objectives is parametrized by the computational analogue of temperature. Similar techniques known as graduated optimization have also been proposed in computer vision [4] and in machine learning, a recent example being [13].

In machine learning, the model complexity is typically controlled through a regularizer. In structural risk minimization, choosing a good regularizer plays an important role in the bias-variance trade-off for model selection [22]. Another virtue of the regularization factor is its influence on the performance of the optimization procedure, as can be seen through continuation methods or *regularization path* techniques [12, 3]. This is formally described by defining a family of objectives functions $f_t := f_{\nu_t}$ with a decreasing sequence ν_t which progressively provide a better approximation of f . The typical goal pursued in this line of work is to combine optimization with model selection, although [12, Section 4.3] also report computational savings. Our use of a continuation method is purely motivated by computational complexity and justified by a rigorous analysis of the quadratic convergence regime of Newton's method as described in detail in the next section.

3 Adaptive Newton Method

3.1 Newton's Method

Assume that we have a μ strongly-convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, which we want to minimize over solutions $\mathbf{x} \in \mathbb{R}^d$. A Newton step defines the increment as

$$\Delta \mathbf{x} = - [\nabla^2 f(\mathbf{x})]^{-1} \nabla f(\mathbf{x}), \quad \mathbf{x} \leftarrow \mathbf{x} + \Delta \mathbf{x} \quad (3)$$

An equivalent way to define Newton increments without the need to invert the Hessian is implicitly as the solution of the linearized optimality condition

$$\nabla f(\mathbf{x} + \Delta \mathbf{x}) \approx \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x}) \Delta \mathbf{x} \stackrel{!}{=} 0 \quad (4)$$

as can be verified by plugging in Eq. (3).

Newton's method converges to the optimal solution $\mathbf{x}^* := \arg \min f(\mathbf{x})$ in a finite number of steps. The speed of convergence is characterized by two distinct phases that depend on the distance to \mathbf{x}^* . The first phase is a *damped phase* with slow convergence while the second phase has quadratic convergence and is triggered when entering a region close to \mathbf{x}^* .

In order to formally characterized this region of quadratic convergence, an important quantity is the *Newton decrement* function [7] defined as

$$\lambda_f := \sqrt{\nabla f^\top [\nabla^2 f]^{-1} \nabla f}, \quad \lambda_f : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}. \quad (5)$$

We will directly make use of this definition in conjunction with an additional requirement that we impose on f , namely that it is *self-concordant* [19]. Self-concordance allows for an elegant, affine-invariant characterization of the quadratic convergence region, as detailed in [7], leading to the sufficient condition that $\lambda_f(\mathbf{x}) \leq \eta$, where $\eta \in (0; \frac{1}{4})$ is a constant whose exact value depends on control parameters for the line search. The self-concordance property restricts the set of functions f , yet it is known that in practice, a similar analysis often optimistically applies to a wider range of functions. For further details, we refer the reader to the discussion in [7, Section 9.6] and the work on logistic regression in [2].

Note that strong convexity implies $\nabla^2 f \succeq \mu \mathbf{I}$ and thus $[\nabla^2 f]^{-1} \preceq \frac{1}{\mu} \mathbf{I}$, so that immediately

$$\lambda_f(\mathbf{x}) \leq \|\nabla f(\mathbf{x})\|_{\frac{1}{\mu} \mathbf{I}} = \frac{1}{\sqrt{\mu}} \|\nabla f(\mathbf{x})\| \implies \|\nabla f(\mathbf{x})\| \stackrel{!}{\leq} \eta \sqrt{\mu} \quad (6)$$

and we arrive at a simple (sufficient) condition on the gradient norm at \mathbf{x} .

Algorithm 1 Basic Newton continuation method: *a priori* (V1) and *data-adaptive* variant (V2).

```

1: given sample  $\mathcal{S}$ , iterations  $T$ 
2: V1: given sequence  $(\mu_t, m_t)$ , V2: given starting point  $(\mu_0, m_0)$ 
3:  $\mathbf{x}_0 \leftarrow \arg \min_{\mathbf{x}} f_{\mu_0, m_0}(\mathbf{x}; \mathcal{S})$ 
4: for  $t = 1, \dots, T$  do
5:   V1: do nothing, V2: compute  $\mu_t$  and  $m_t$ 
6:   compute Newton increment  $\Delta \mathbf{x}$  for  $f_t(\mathbf{x}) := f_{\mu_t, m_t}(\mathbf{x}_{t-1}, \mathcal{S})$ 
7:    $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} + \Delta \mathbf{x}$ 
8: end for

```

3.2 Continuation Method

We study the 2-parametric family of regularized empirical risk functions that are defined via smooth and convex, non-negative loss functions ϕ^z and relative to a full sample $\mathcal{S}^* = (z_1, \dots, z_N)$. We make use of the definitions in Eq. (1) and we think of $f_\nu^{\mathcal{S}}$ as being indexed by ν as well as $n = |\mathcal{S}|$, where $\mathcal{S} = (z_1, \dots, z_n)$ consists of the first n samples of \mathcal{S}^* . The canonical regularizer we consider is $\Omega(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|^2$ and we will sometimes utilize the commonly used heuristics of choosing $\mu \propto 1/m$ to focus on a simpler 1-parametric family.

We want to implement the abstract procedure described in Algorithm 1, where we either pre-generate a sequence of problems f_t or, alternatively, construct (μ_t, m_t) greedily in a data-adaptive manner. The key condition for making this work as expected (by design) is that we are able to establish the following condition

$$\lambda_{f_t}(\mathbf{x}_{t-1}^*) \leq \eta \quad \text{or, more conservatively,} \quad \|\nabla f_t(\mathbf{x}_{t-1}^*)\| \leq \eta \sqrt{\mu_t}, \quad (7)$$

which will assure (under appropriate assumptions, e.g. self-concordance) that the minimizer of the previous optimization problem will provide a starting point that is within the quadratic convergence region of the subsequent optimization problem, yielding a proper "hand-over" of the solution as illustrated in Figure 1.

3.3 Reducing Regularization Strength

Let us first assume that we fix \mathcal{S} and that $\mu := \mu_{t-1}$. We are seeking for the range of possible $\nu := \mu_t \leq \mu$ for which we can guarantee the condition in Eq. (7).

Lemma 1. Let $\Omega(\cdot) = \frac{1}{2} \|\cdot\|^2$, $\mathbf{x}_\mu^* := \arg \min_{\mathbf{x}} f_\mu(\mathbf{x})$, and $B_\mu := \frac{\eta^2}{\mu \|\mathbf{x}_\mu^*\|^2}$. For any ν such that

$$\mu \geq \nu \geq \mu \left(1 - \frac{1}{2} \left(\sqrt{B_\mu^2 + 4B_\mu} - B_\mu \right) \right) \implies \lambda_{f_\nu}(\mathbf{x}_\mu^*) \leq \eta \quad (8)$$

Proof. By the first order optimality condition for \mathbf{x}_μ^* , we have that $\nabla f_0(\mathbf{x}_\mu^*) = -\mu \mathbf{x}_\mu^*$, hence

$$\nabla f_\nu(\mathbf{x}_\mu^*) = \nabla f_0(\mathbf{x}_\mu^*) + \nu \mathbf{x}_\mu^* = (\nu - \mu) \mathbf{x}_\mu^* \iff \|\nabla f_\nu(\mathbf{x}_\mu^*)\| = (\mu - \nu) \|\mathbf{x}_\mu^*\|. \quad (9)$$

By definition we have

$$\lambda_{f_\nu}(\mathbf{x}_\mu^*) = \frac{1}{\sqrt{\nu}} \|\nabla f_\nu(\mathbf{x}_\mu^*)\| \stackrel{!}{\leq} \eta \iff \|\mathbf{x}_\mu^*\| \stackrel{!}{\leq} \frac{\eta \sqrt{\nu}}{(\mu - \nu)} \iff \nu^2 - (2 + B_\mu) \mu \nu + \mu^2 \stackrel{!}{\leq} 0 \quad (10)$$

Solving the quadratic equation for ν yields the claim. \square

Corollary 2. Assume that $\phi(\mathbf{0}, z) \leq \Phi (\forall z)$. Lemma 1 remains valid with $B = \frac{\eta^2}{2\Phi} \geq B_\mu$.

Proof. One can bound $\|\mathbf{x}_\mu^*\|$ easily through the following argument, exploiting non-negativity of ϕ

$$\Phi \geq f_0(\mathbf{0}) = f_\mu(\mathbf{0}) \geq f_\mu(\mathbf{x}_\mu^*) = f_0(\mathbf{x}_\mu^*) + \frac{\mu}{2} \|\mathbf{x}_\mu^*\|^2 \geq \frac{\mu}{2} \|\mathbf{x}_\mu^*\|^2 \implies \|\mathbf{x}_\mu^*\|^2 \leq \frac{2\Phi}{\mu}. \quad (11)$$

\square

Note that Corollary 2 gives an *a priori* rate guarantee, which only depends on the easily computable Φ . This is a strong argument in favor of a continuation method, yet may not result in the most efficient schedules for reducing μ . We will revisit this issue in Section 3.5.

3.4 Increasing Sample Size

We now generalize our analysis to the case, where we also increase the sample size. The basic challenge is that we need to upper bound the gradient norm contributions coming from new data points. Intuitively this depends on the generalization capability of the current iterate.

Lemma 3. *Assume f has Lipschitz continuous gradients with constant L . Let \mathcal{S} be a given n -sample, which we split into $\mathcal{S}_0 = (z_1, \dots, z_m)$ and $\mathcal{S}_1 = (z_{m+1}, \dots, z_n)$, where $m \leq n$ is arbitrary. Define $\mathbf{x}^* := \arg \min_{\mathbf{x}} f^{\mathcal{S}_0}(\mathbf{x})$. With high probability over \mathcal{S}_0 it holds that*

$$\mathbf{E}_{\mathcal{S}_1} \|\nabla f^{\mathcal{S}}(\mathbf{x}^*)\|^2 \leq \frac{2L(n-m)}{n} \mathcal{H}(m) \quad (12)$$

Proof. See Appendix. □

Theorem 4. *Under the same conditions as Lemma 3 and for arbitrary $\nu \leq \mu$ and $\beta \in (0; \infty)$,*

$$\mathbf{E}_{\mathcal{S}_1} \|\nabla f_{\nu}^{\mathcal{S}}(\mathbf{x}^*)\|^2 \leq (1 + \beta) (\nu - \mu)^2 \|\mathbf{x}^*\|^2 + (1 + \beta^{-1}) \left(\frac{n-m}{n} \right)^3 2L\mathcal{H}(m)$$

Proof. See Appendix. □

The theorem directly implies that we can construct a sequence of objective functions with sample size increasing at a geometric rate.

Corollary 5. *There exists an $\alpha^* \in [0; 1)$ such that for all $\alpha \in [\alpha^*; 1]$, with $m/\alpha \in \mathbb{Z}$, the following holds for $\nu := \alpha\mu$ and $\mathcal{S} = (z_1, \dots, z_{m/\alpha})$,*

$$\|\nabla f_{\nu}^{\mathcal{S}}(\mathbf{x}^*)\| \leq \eta\sqrt{m}, \quad (13)$$

where α^* can be explicitly computed as the root of a third order polynomial.

Proof. See Appendix. □

3.5 Data-Adaptive Algorithm

The above analysis is largely data-independent and just involves a few constants: L , Φ , proportionality constants involving \mathcal{H} and $\mu \propto 1/m$. As such, it leads to geometric rates that can be quite conservative. We will now show a data-adaptive strategy that – up to small approximation errors – maximally increases the sample size, and, equivalently, maximally decreases the regularization strength, within the desired range. First of all, note that we can easily compute the gradient on an increased sample. Denote $\mathcal{S} = (z_1, \dots, z_n)$ and $\mathcal{S}_0 = (z_1, \dots, z_m)$, $m \leq n$. Define $\mathbf{x}^* = \arg \min_{\mathbf{x}} f_{\mu}^{\mathcal{S}_0}(\mathbf{x})$,

$$\nabla f_{\nu}^{\mathcal{S}}(\mathbf{x}^*) = \frac{1}{n} \sum_{k=m+1}^n \nabla \phi^{z_k}(\mathbf{x}^*) + \nu \left(1 - \frac{\mu m}{\nu n} \right) \mathbf{x}^*. \quad (14)$$

Note that if $\mu/\nu = n/m$, then the second term vanishes. As we need to compute the gradient anyway as part of the Newton iteration, we get it for free. We now could approximate

$$\lambda_{f_{\nu}^{\mathcal{S}}}(\mathbf{x}^*) \leq \frac{1}{\sqrt{\nu}} \|\nabla f_{\nu,n}(\mathbf{x}^*)\|, \quad (15)$$

but this results in bounds that are not very tight and hence underestimate the possible rate. We thus propose a tighter bound based on a Taylor approximation of the inverse Hessian at the previous iteration, something that we can compute with a little bit of extra computational effort.² Let $\nu \leq \mu$ and define $\mathbf{H} := \nabla^2 f_{\mu}^{\mathcal{S}_0}(\mathbf{x}^*)$

$$[\mathbf{H} - (\mu + \nu)\mathbf{I}]^{-1} = \mathbf{H}^{-1} + (\mu - \nu)\mathbf{H}^{-2} + \mathbf{O}(\mu^2) \quad (16)$$

²Computing a Newton update is usually done via LU decomposition, which is cheaper.

Algorithm 2 Data-adaptive Newton’s method - DYNANEWTON

```

1: Given sample  $\mathcal{S}$  and  $(\mu_0, m_0)$ , define  $\mathcal{S}_0 := \mathcal{S}_{1:m_0}$ 
2:  $\mathbf{x}_0 \leftarrow \arg \min_{f_{\mu_0}^{\mathcal{S}_0}}$ 
3: for  $t = 1, \dots, T$  do
4:   Find the smallest  $\alpha$  such that
       $\lambda_{f_{\nu}^{\mathcal{S}}}(\mathbf{x}^*) \leq \eta$ , where  $f^{\alpha} := f_{\nu}^{\mathcal{S}'}$ ,  $\nu := \alpha\mu$ ,  $\mathcal{S}' := \mathcal{S}_{1:n}$ ,  $n := m_{t-1}/\alpha$ .
5:    $(\mu_t, m_t) \leftarrow (\nu, n)$ 
6:   Compute Newton increment  $\Delta \mathbf{x}$  for  $f^{\alpha}$  at  $\mathbf{x}_{t-1}$ 
7:    $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} + \Delta \mathbf{x}$ 
8: end for

```

and thus with the additional assumption that $\mathbf{H} \approx \nabla^2 f_{\mu}^{\mathcal{S}}(\mathbf{x}^*)$,

$$[\lambda_{f_{\nu}^{\mathcal{S}}}(\mathbf{x}^*)]^2 \approx \nabla f_{\nu}^{\mathcal{S}}(\mathbf{x}^*)^{\top} \mathbf{H}^{-1} \nabla f_{\nu}^{\mathcal{S}}(\mathbf{x}^*) + (\mu - \nu) \|\mathbf{H}^{-1} \nabla f_{\nu}^{\mathcal{S}}(\mathbf{x}^*)\|^2. \quad (17)$$

For instance, in the setting $\nu \propto 1/n$, it is straightforward to (numerically) find the maximal n – or equivalently minimal ν – such that the above approximation is $< \eta^2$. In our experiments, we have found this approximation to be correct up to a few percent relative error.

The complete algorithm is summarized in Algorithm 2. In order to find an α that fulfils Eq. (22), we need to compute an estimate of $\|\nabla f_{\mu}^{\mathcal{S}}(\mathbf{x}^*)\|$. We solve this problem by first assuming that \mathbf{x}_t is a good approximation of the solution to f_t and then computing the estimate from the samples $(z_{m+1}, \dots, z_{m+k})$. We can avoid the first approximation at the cost of making 5 – 6 Newton iterations instead of just 1, but we have found this not to be necessary in any one of our experiments. Further experimental results are provided in the appendix.

4 Experiments

Datasets We apply ℓ_2 -regularized logistic regression on a set of 4 datasets of varying size and dimensionality summarized in Table 1. We use 90% of the data points as the training set and the remaining 10% as test set.

DATASET	SIZE	NUMBER OF FEATURES
A9A	32561	123
W8A	49749	300
COVTYPE.BINARY	581012	54
SUSY	5000000	18

Table 1: *Details of the datasets used in our experiments.*

Comparison to standard baselines We compare DYNANEWTON (cf. Algorithm 2) to Newton’s method and – as a competitive SGD variant – to SAGA. We used a step size of $\sim \frac{1}{L}$ for SAGA as suggested in previous work [10, 14]. The results presented in Figure 2 show significant speeds-ups compared to Newton’s method. DYNANEWTON also outperforms SAGA and gets a very accurate solution after less than 6 epochs on all datasets. We also evaluated the solution quality on the expected risk and provide the complete results of this evaluation in the Appendix (see Figure 6). In order to relate convergence on the empirical and expected risks, we plotted a horizontal dotted line that shows the iteration at which DYNANEWTON reached convergence on the test set. This demonstrates that DYNANEWTON also achieves significant gains in terms of test error.

As the cost per iteration is typically higher for Newton’s method on a single machine, we also present a comparison in terms of running time in the Appendix (see Figure 5). Although the gains are more moderate when measuring time, it is worth pointing out that Newton is inherently much easier to parallelize as discussed in [9] and further gains are thus to be expected in a distributed setting.

Increment factor test We evaluate the influence of α on the convergence of the algorithm without data-adaptivity. As can be seen in Figure 4 a small value of α leads to faster convergence while it might also make the algorithm diverge if chosen too aggressively small. In contrast, the data-adaptive approach adapts the value of α and yields a stable convergence behavior on all datasets.

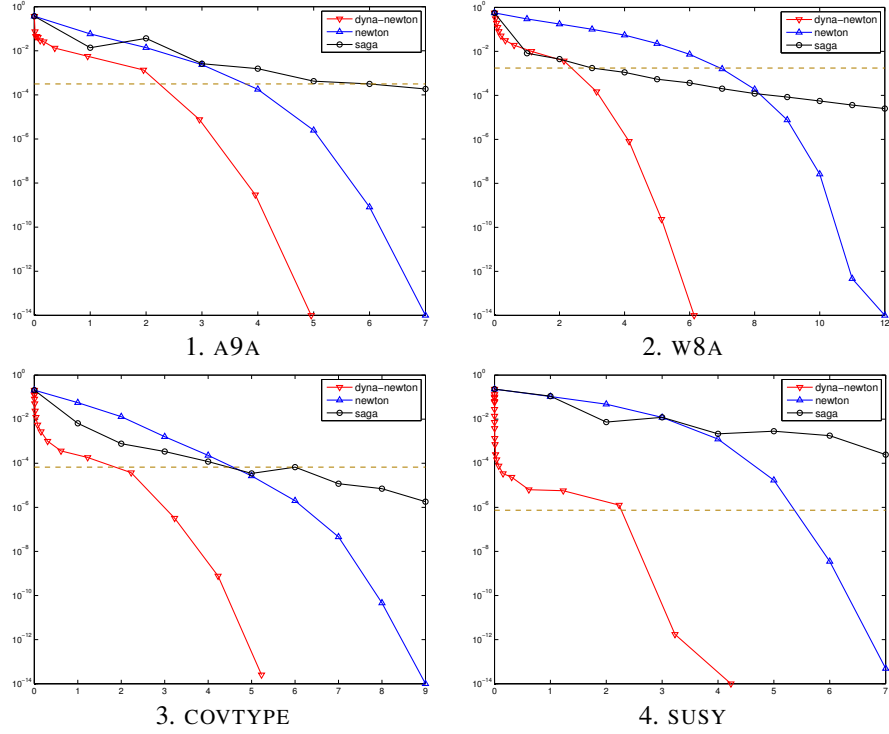


Figure 2: *Suboptimality on the empirical risk vs effective number of epochs.* The plots show the suboptimality of the empirical risk on the full data set \mathcal{S}^* with $\nu = 1/N$. The horizontal axis represents the number of *effective epochs*, i.e. number of passes over the whole training data set. The dotted horizontal line refers to statistical accuracy (explained in the main text).

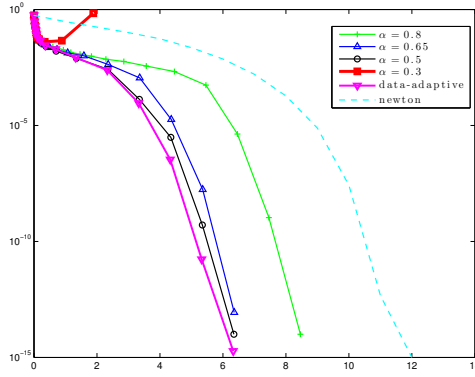


Figure 3: *Increment Factor test on the w8A dataset.* The plots shows the suboptimality of the empirical risk on the full data set \mathcal{S}^* with $\nu = 1/N$. In these experiments we used the increment factor $\alpha = m_{t-1}/m_t$. It is clear that the smallest α value causes leaving the region of quadratic convergence and hence leads to divergence. This behavior is excluded by the data-adaptive approach.

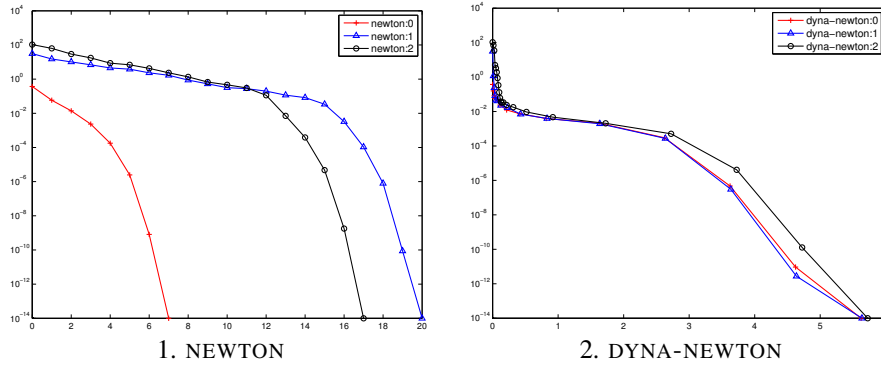


Figure 4: *Robustness against the initialisation point on the A9A dataset.* We here show the suboptimality on empirical risk against time for various initialization points: in red, $x_0 = \vec{0}$, in blue $x_0 = \vec{3}$ and in black $x_0 = \vec{10}$.

We observed empirically that the data-adaptive method chooses a value of $\alpha \approx \frac{1}{2}$ thus explaining why the non-adaptive approach with $\alpha = \frac{1}{2}$ achieves a similar - but slightly inferior - performance.

Importance of the initialization point We investigate the role of the initialization point on the convergence of Newton’s method and DYNANEWTON. The results shown in Figure 4 demonstrate that bad initialization points (i.e. far away from the optimum) significantly slow down the convergence of Newton’s method as they require more steps to get inside the ball of quadratic convergence. By comparison, a poor initialization does not significantly affect DYNANEWTON.

5 Conclusion

We proposed a continuation method variant of Newton’s method that dynamically adapts the sample size and the regularization strength such that each subproblem provides a starting point that is within the quadratic convergence region of the subsequent optimization problem. We provided a theoretical analysis that characterizes the conditions required for achieving a proper hand-over and also devised a data-adaptive strategy of discretizing the solution path.

Our empirical results demonstrate significant speed-ups across a wide range of datasets both in terms of empirical as well as expected risk. In particular the speed of convergence on the latter is quite

remarkable, often getting close to optimal solutions in about 2 effective epochs. All results seem to be in good agreement with our theory and its predictions.

In the appendix, we provide further empirical evidence that shows that our results extend to non-exact Newton methods such as L-BFGS. This is of special interest for training large deep networks for which a distributed implementation of L-BFGS has been shown to provide significant speed-ups in terms of training time [9]. While our analysis does not provide any guarantees for this case, it seems that the proposed continuation method has merits beyond the results presented in this paper. A better understanding of the behavior of quasi-Newton methods is the topic of future work.

Acknowledgments

We would like to thank Aryan Mokhtari for helpful discussions on Newton’s method.

References

- [1] E. L. Allgower and K. Georg. *Numerical continuation methods: an introduction*, volume 13. Springer Science & Business Media, 2012.
- [2] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.
- [3] F. R. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. *Advances in neural information processing systems*, 17:73–80, 2005.
- [4] A. Blake and A. Zisserman. *Visual reconstruction*, volume 2. MIT press Cambridge, 1987.
- [5] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.
- [6] O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.
- [7] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [8] H. Daneshmand, A. Lucchi, and T. Hofmann. Starting small - learning with adaptive sample sizes. In *International Conference on Machine Learning*, 2016.
- [9] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, pages 1223–1231, 2012.
- [10] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.
- [11] J. E. Dennis, Jr and J. J. Moré. Quasi-newton methods, motivation and theory. *SIAM review*, 19(1):46–89, 1977.
- [12] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *The Journal of Machine Learning Research*, 5:1391–1415, 2004.
- [13] E. Hazan, K. Y. Levy, and S. Shalev-Swartz. On graduated optimization for stochastic non-convex problems. *arXiv preprint arXiv:1503.03712*, 2015.
- [14] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems* 28, pages 2296–2304, 2015.
- [15] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.
- [16] S. Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of statistical physics*, 34(5-6):975–986, 1984.
- [17] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.
- [18] A. Mokhtari and A. Ribeiro. Res: Regularized stochastic bfgs algorithm. *Signal Processing, IEEE Transactions on*, 62(23):6089–6104, 2014.

- [19] Y. Nesterov, A. Nemirovskii, and Y. Ye. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.
- [20] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*, volume 30. SIAM, 1970.
- [21] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.
- [22] V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.
- [23] J. Wang, H. Wang, and N. Srebro. Reducing runtime by recycling samples. *arXiv preprint arXiv:1602.02136*, 2016.

A Appendix

A.1 Proofs

Lemma 3

Proof. Using classical concentration inequality from statistical learning theory we get (see also [8]) with high probability over \mathcal{S}_0

$$\mathbb{E}_{\mathcal{S}_1} \left[f^{\mathcal{S}}(\mathbf{x}^*) - \min_{\mathbf{x}} f^{\mathcal{S}}(\mathbf{x}) \right] \leq \frac{n-m}{n} \mathcal{H}(m), \quad (18)$$

which bounds the suboptimality of the minimizer of the smaller m -sample on the larger n -sample. Furthermore, by virtue of smoothness:

$$\|\nabla f^{\mathcal{S}}(\mathbf{x}^*)\|^2 \leq 2L \left[f^{\mathcal{S}}(\mathbf{x}^*) - \min_{\mathbf{x}} f^{\mathcal{S}}(\mathbf{x}) \right] \quad (19)$$

Putting both inequalities together yields the claim. \square

Theorem 4

Proof.

$$\begin{aligned} \nabla f_{\nu}^{\mathcal{S}}(\mathbf{x}^*) &= \frac{m}{n} \nabla f_0^{\mathcal{S}_0}(\mathbf{x}^*) + \frac{n-m}{n} \nabla f_0^{\mathcal{S}_1}(\mathbf{x}^*) + \nu \mathbf{x}^* \\ &= -\frac{m}{n} \mu \mathbf{x}^* + \frac{n-m}{n} \nabla f_{\mu}^{\mathcal{S}_1}(\mathbf{x}^*) - \frac{n-m}{n} \mu \mathbf{x}^* + \nu \mathbf{x}^* \\ &= (\nu - \mu) \mathbf{x}^* + \frac{n-m}{n} \nabla f_{\mu}^{\mathcal{S}_1}(\mathbf{x}^*) = (\nu - \mu) \mathbf{x}^* + \frac{n-m}{n} \nabla f_{\mu}^{\mathcal{S}}(\mathbf{x}^*). \end{aligned} \quad (20)$$

Here we have repeatedly exploited the first order optimality condition of \mathbf{x}^* , i.e. $\nabla f_{\mu}^{\mathcal{S}_1}(\mathbf{x}^*) = 0$. Now, taking squared norms on both sides, we apply the generalized parallelogram identity $\|a + b\|^2 \leq (1 + \beta)\|a\|^2 + (1 + \beta^{-1})\|b\|^2$ with $\beta \in (0; \infty)$.

$$\|\nabla f_{\nu}^{\mathcal{S}}(\mathbf{x}^*)\|^2 \leq (1 + \beta) (\nu - \mu)^2 \|\mathbf{x}^*\|^2 + (1 + \beta^{-1}) \left(\frac{n-m}{n} \right)^2 \|\nabla f_{\mu}^{\mathcal{S}}(\mathbf{x}^*)\|^2 \quad (21)$$

Taking expectation with regard to \mathcal{S}' on both sides and applying Lemma 3 concludes the proof. \square

Corollary 5

Proof. For concreteness set $\beta = 1$. We want to chose α such that

$$\mu^2 \|\mathbf{x}^*\|^2 (1 - \alpha)^2 + 2L\mathcal{H}(m) (1 - \alpha)^3 \leq \frac{\mu\eta^2}{2} \quad (22)$$

For $\alpha = 1$, the left hand side is 0, while the right hand side is strictly positive. As the derivative of the left hand side is finite at 1, we can solve for $\alpha < 1$ (such that $\alpha \geq 0$) as claimed. \square

A.2 Running time

We compare the running time of DYNANEWTON against other baselines in Figure 5. Although we see more moderate gains in comparison to SAGA in terms of computational time (especially in early iterations), it is worth pointing out that Newton is inherently much easier to parallelize as discussed in [9] and further gains are thus to be expected in a distributed setting.

A.3 Expected risk

We present the results in terms of expected error in Figure 6. This largely confirms the results obtained on the training set. Although all methods achieve convergence after less than 3 epochs, DYNANEWTON achieves even faster convergence than Newton and SAGA.

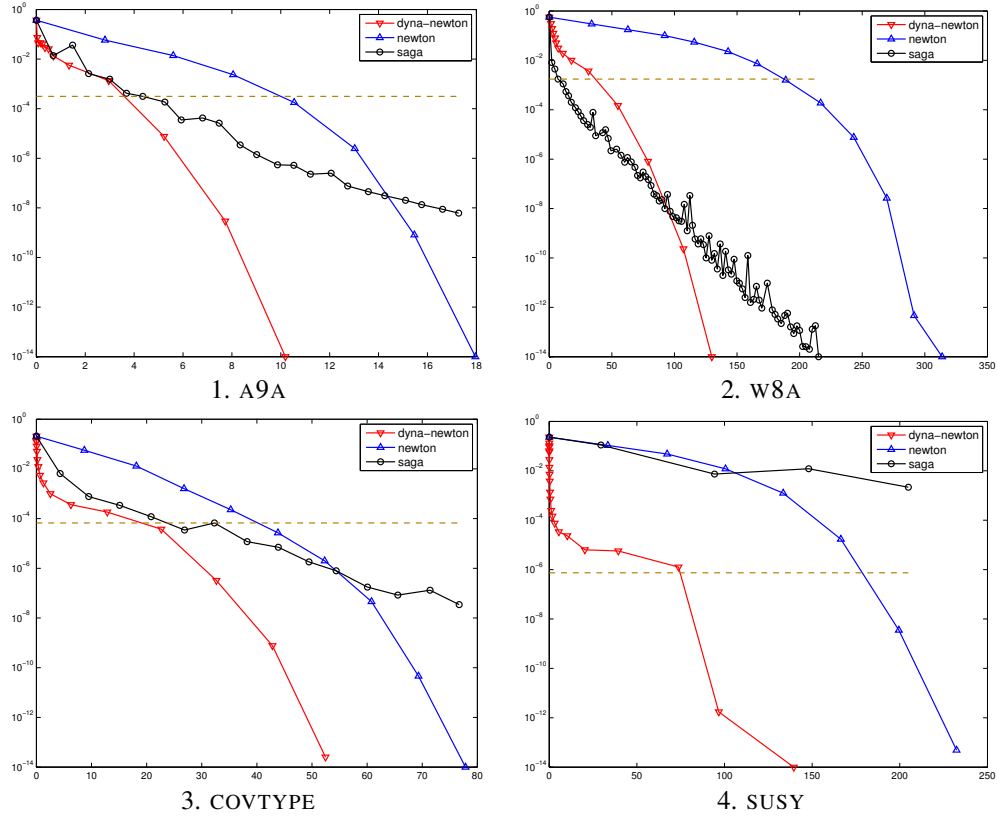


Figure 5: *Suboptimality on the empirical risk vs time.* The vertical axis shows the suboptimality of the empirical risk, i.e. $f_{\nu,n}(x^t) - f_{\nu,n}^*$, where $\nu = 1/n$. The horizontal axis represents run time in seconds. The horizontal dotted line corresponds to the iteration at which convergence is reached on the expected risk (see details in the main text).

A.4 Approximation

The analysis we developed assumes that we reach $x_\mu^* := \arg \min_x f_\mu(x)$ before switching to ν . We empirically check the robustness to an approximate solution \hat{x}_μ obtained by performing a single Newton step instead of 6 steps, which guarantees convergence up to numerical precision (see Section 9.5 in [7]). As shown in Figure 7, the impact of the resulting approximate solution is almost negligible.

A.5 BFGS

One shortcoming of Newton’s method is that it requires solving a linear equation system involving the Hessian matrix, which may be impractical for large and high-dimensional datasets. Approximate variants known as quasi-Newton methods [11] have thus been developed, such as the popular BFGS or its limited memory version L-BFGS [17]. Quasi-Newton methods such as BFGS do not require computing the Hessian matrix but instead construct a quadratic model of the objective function by successive evaluations of the gradient. There seems to be a gap between theoretically guaranteed convergence rates and the empirically observed effectiveness of BFGS, in particular on ill-conditioned problems [18] and for non-convex problems [9].

We used the inspiration provided by our continuation method approach to develop a variant of L-BFGS that like DYNALBFGS, adaptively changes the sample size and the regularizer. We name this approach DYNALBFGS. The pseudo-code of this method is the same as Algorithm 2 except that the Newton increment is computed from the approximate L-BFGS Hessian matrix instead of the true Hessian. We evaluate the performance of DYNALBFGS for the task of ℓ_2 -regularized logistic re-

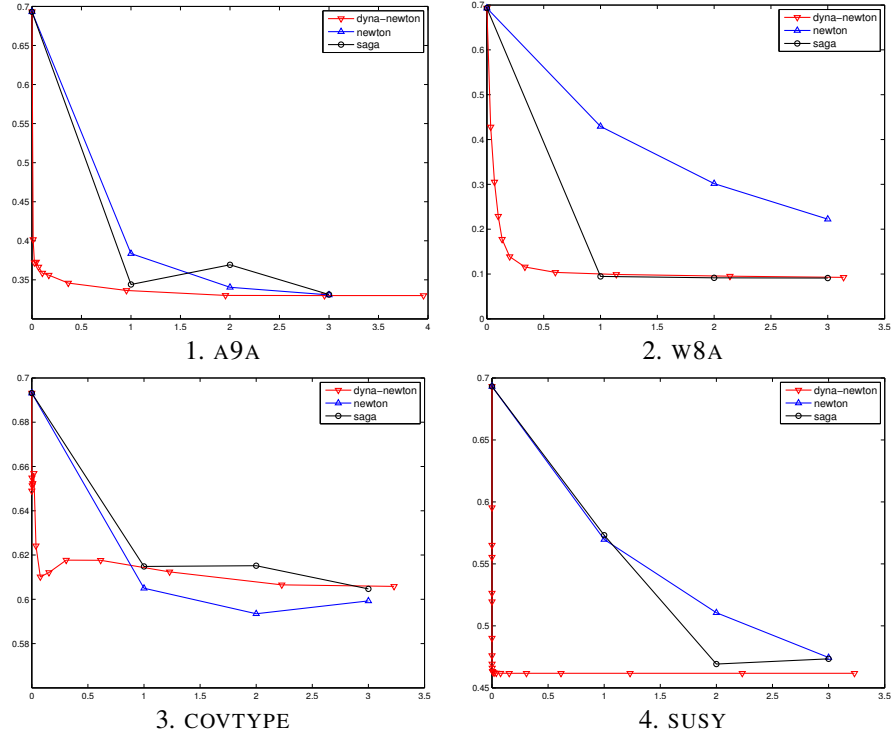
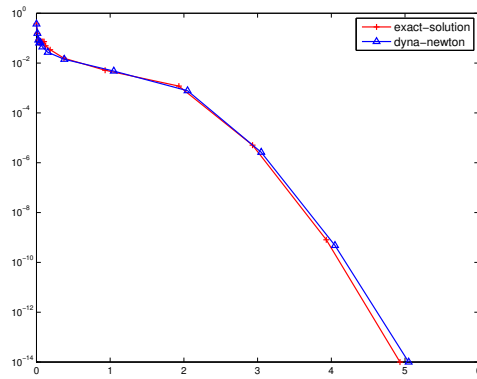
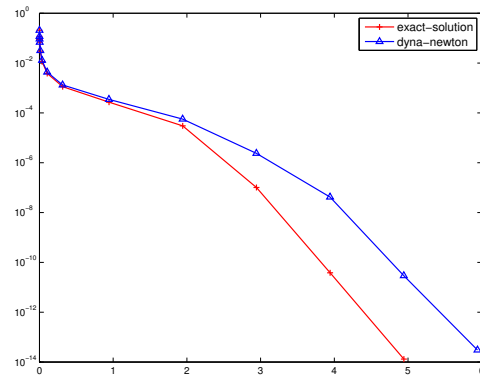


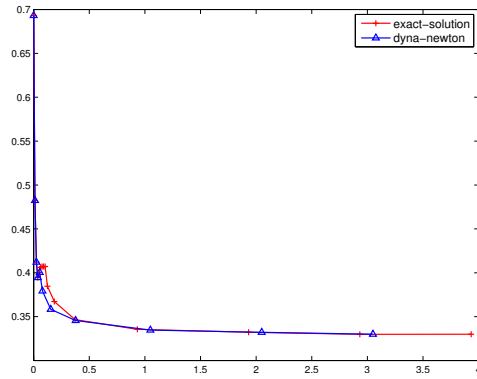
Figure 6: *Suboptimality on the test set vs number of epochs.* The vertical axis shows the suboptimality of the expected risk, i.e. $\phi_{\mathcal{T}}(\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{T}} \phi_{\mathbf{z}}(\mathbf{x}) / |\mathcal{T}|$, where \mathcal{T} is the test set. The horizontal axis represents the number of epochs.



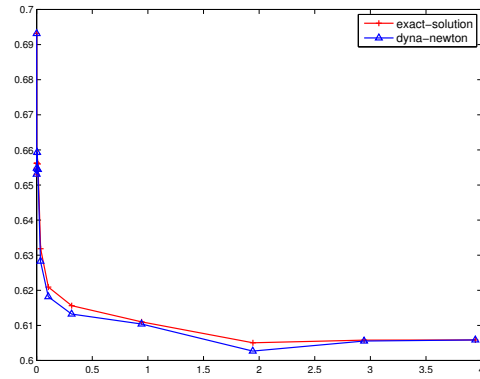
1. Empirical suboptimality on A9A



2. Empirical suboptimality on COVTYPE



3. Test error on A9A



4. Test error on COVTYPE

Figure 7: *Effect of the approximate solution obtained by DYNANEWTON.* We here check the effect of the approximate solution obtained by DYNANEWTON by iterating for 1 epoch against the exact solution obtained by iterating for 6 epochs.

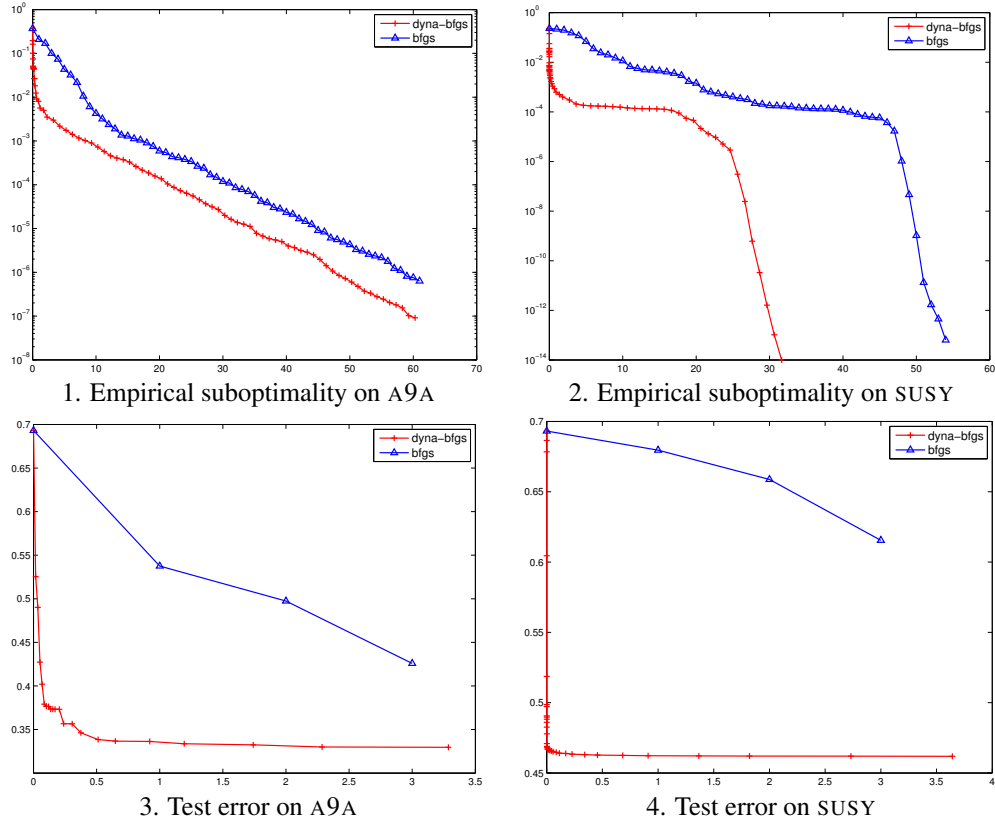
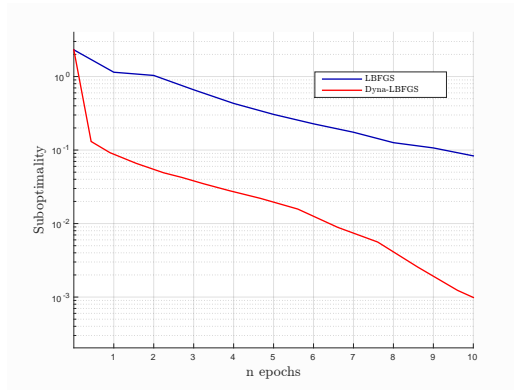
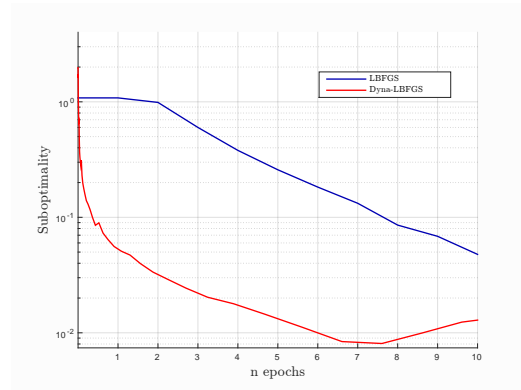


Figure 8: *Comparison of BFGS vs DYNALBFGS.* We here provide the empirical suboptimality as well as the test error on the A9A and SUSY datasets. We would like to point out that the range of the y-axis between the top and bottom row is different as convergence on the test set was typically observed after 1 or 2 epochs.

gression on the two datasets described in the main paper. The results shown in Figure 8 demonstrate significant gains compared to L-BFGS. We also investigate the performance of DYNALBFGS on training a convolutional neural network consisting of two convolutional and pooling layers with one fully-connected layer. We include results on the standard MNIST dataset in Figure 9. Although our analysis does not extend to non-convex functions, we nevertheless still observe significant gains.



1. Training error



2. Test error

Figure 9: *Comparison of BFGS vs DYNABFGS for training neural networks.* We here show the performance of DYNABFGS to train a standard convolutional neural network on the MNIST dataset.